

CS:4420 Artificial Intelligence

Spring 2017

Uncertainty

Cesare Tinelli

The University of Iowa

Copyright 2004–17, Cesare Tinelli and Stuart Russell ^a

^a These notes were originally developed by Stuart Russell and are used with permission. They are copyrighted material and may not be used in other course settings outside of the University of Iowa in their current or modified form without the express written consent of the copyright holders.

Readings

- Chap. 13 of [Russell and Norvig, 2012]

Logic and Uncertainty

Major problem with logical-agent approaches:

Agents almost never have access to the whole truth about their environments.

- Very often, even in simple worlds, there are important questions for which there is no boolean answer.
- In that case, an agent must reason under **uncertainty**.
- Uncertainty also arises because of an agent's incomplete or incorrect understanding of its environment.

Uncertainty

Let action A_t = “leave for airport t minutes before flight”

Will A_t get me there on time?

Problems

- partial observability (road state, other drivers' plans, etc.)
- noisy sensors (unreliable traffic reports)
- uncertainty in action outcomes (flat tire, etc.)
- immense complexity of modeling and predicting traffic

Uncertainty

Let action A_t = “leave for airport t minutes before flight”

Will A_t get me there on time?

Hence a purely logical approach either

1. risks falsehood (“ A_{25} will get me there on time”), or
2. leads to conclusions that are too weak for decision making (“ A_{25} will get me there on time if there’s no accident on the way, it doesn’t rain, my tires remain intact, . . .”)

(A_{1440} might reasonably be said to get me there on time but I’d have to stay overnight in the airport . . .)

Reasoning under Uncertainty

A rational agent is one that makes rational decisions (in order to maximize its performance measure)

A rational decision depends on

- the **relative importance** of various goals
- the **likelihood** they will be achieved
- the **degree** to which they will be achieved

Handling Uncertain Knowledge

Reasons FOL-based approaches fail to cope with domains like, for instance, medical diagnosis:

- **Laziness:** too much work to write complete axioms, or too hard to work with the enormous sentences that result.
- **Theoretical Ignorance:** The available knowledge of the domain is incomplete.
- **Practical Ignorance:** The theoretical knowledge of the domain is complete but some evidential facts are missing.

Degrees of Belief

- In several real-world domains the agent's knowledge can only provide a **degree of belief** in the relevant sentences
- The agent cannot say whether a sentence is true, but only that it is true $x\%$ of the times
- The main tool for handling degrees of belief is **Probability Theory**
- The use of probability summarizes the uncertainty that stems from our laziness or ignorance about the domain

Probability Theory

Probability Theory makes the same ontological commitments as First-order Logic:

Every sentence φ is either true or false

The degree of belief that φ is true is a number P between 0 and 1

$P(\varphi) = 1$ \longrightarrow φ is certainly true

$P(\varphi) = 0$ \longrightarrow φ is certainly not true.

$P(\varphi) = 0.65$ \longrightarrow φ is true with a 65% chance

Probability of Facts

Let A be a propositional variable, a symbol denoting a proposition that is either true or false

$P(A)$ denotes the probability that A is true *in the absence of any other information*

Similarly,

$P(\neg A)$ = probability that A is false

$P(A \wedge B)$ = probability that both A and B are true

$P(A \vee B)$ = probability that either A or B (or both) are true

Examples:

$P(\neg \text{Blonde})$ $P(\text{Blonde} \wedge \text{BlueEyed})$ $P(\text{Blonde} \vee \text{BlueEyed})$

where $\text{Blonde}/\text{BlueEyed}$ denotes that a given person is blonde/blue-eyed

Subjective/Bayesian Probability

Probabilities relate propositions to one's own state of knowledge

E.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

Probabilities of propositions change with new evidence:

E.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Note: This is analogous to logical entailment status $KB \models \alpha$ (which changes with more knowledge), not truth

Conditional/Unconditional Probability

$P(A)$ is the **unconditional (or prior) probability** of fact A

An agent can use the unconditional probability of A to reason about A in the absence of further information

If further evidence B becomes available, the agent must use the **conditional (or posterior) probability**:

$$P(A | B)$$

the probability of A given that (all) the agent knows (is) B

Note: $P(A)$ can be thought as the conditional probability of A with respect to the empty evidence: $P(A) = P(A |)$

Conditional Probabilities

The probability of a fact may change as the agent acquires more, or different, information:

1. $P(\textit{Blonde})$
2. $P(\textit{Blonde} \mid \textit{Swedish})$
3. $P(\textit{Blonde} \mid \textit{Kenian})$
4. $P(\textit{Blonde} \mid \textit{Kenian} \wedge \neg \textit{EuroDescent})$

1. If we know nothing about a person, the probability that s/he is blonde equals a certain value, say **0.2**
2. If we know that a person is Swedish the probability that s/he is blonde is much higher, say **0.9**
3. If we know that the person is Kenyan, the probability s/he is blonde much lower, say **0.00003**
4. If we know that the person is Kenyan and not of European descent, the probability s/he is blonde is **0**

The Axioms of Probability

Probability Theory is governed by the following **axioms**:

1. All probabilities are real values between **0** and **1**:

$$\text{for all } \varphi, \quad 0 \leq P(\varphi) \leq 1$$

2. Valid propositions have probability **1**

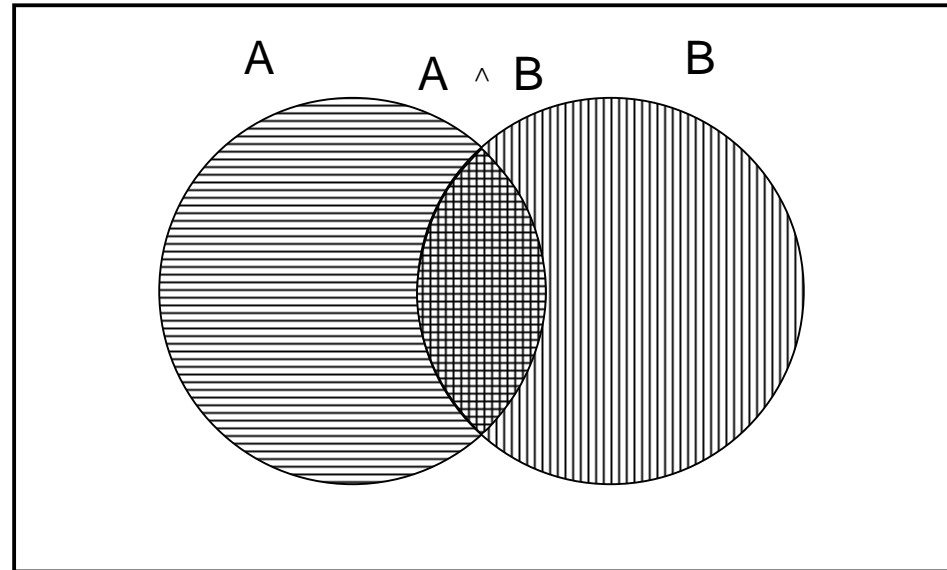
$$P(\mathbf{True}) = P(\alpha \vee \neg\alpha) = 1$$

3. The probability of disjunction is defined as follows:

$$P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$$

Understanding Axiom 3

True



$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Conditional Probabilities

Conditional probabilities are defined in terms of unconditional ones.

Whenever $P(B) > 0$,

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

The same definition can be equivalently expressed as the **product rule**:

$$\begin{aligned} P(A \wedge B) &= P(A | B)P(B) \\ &= P(B | A)P(A) \end{aligned}$$

A and B are **independent** iff $P(A | B) = P(A)$ (or $P(B | A) = P(B)$, or $P(A \wedge B) = P(A)P(B)$)

Random Variable

A **random variable** is a variable ranging over a certain domain of values

It is

- **discrete** if it ranges over a discrete (that is, countable) domain
- **continuous** if it ranges over the real numbers

We will only consider discrete random variables with finite domains

Note: Propositional variables can be seen as random variables over the Boolean domain

Random Variables

Variable	Domain
<i>Age</i>	$\{1, 2, \dots, 120\}$
<i>Weather</i>	$\{sunny, dry, cloudy, rain, snow\}$
<i>Size</i>	$\{small, medium, large\}$
<i>Blonde</i>	$\{true, false\}$

The probability that a random variable X has value val is written as

$$P(X = val)$$

Note 1: $P(A = true)$ is written shortly as $P(a)$ while $P(A = false)$ is written as $P(\neg a)$

Note 2: Traditionally, in Probability Theory variables are capitalized and constant values are not

Probability Distribution

If X is a random variable, we use the bold case $\mathbf{P}(X)$ to denote a *vector* of values for the probabilities of each individual element that X can take.

Example:

$$P(\textit{Weather} = \textit{sunny}) = 0.6$$

$$P(\textit{Weather} = \textit{rain}) = 0.2$$

$$P(\textit{Weather} = \textit{cloudy}) = 0.18$$

$$P(\textit{Weather} = \textit{snow}) = 0.02$$

Then $\mathbf{P}(\textit{Weather}) = \langle 0.6, 0.2, 0.18, 0.02 \rangle$
(the value order of “sunny”, “rain”, “cloudy”, “snow” is assumed).

$\mathbf{P}(\textit{Weather})$ is called a **probability distribution** for the random variable $\textit{Weather}$

Joint Probability Distribution

If X_1, \dots, X_n are random variables,

$$\mathbf{P}(X_1, \dots, X_n)$$

denotes their **joint probability distribution (JPD)**, an n -dimensional matrix specifying the probability of every possible combination of values for X_1, \dots, X_n

Example

Sky : {*sunny, cloudy, rain, snow*}

Wind : {*true, false*}

$\mathbf{P}(\textit{Wind}, \textit{Sky}) =$

	<i>sunny</i>	<i>cloudy</i>	<i>rain</i>	<i>snow</i>
<i>true</i>	0.30	0.15	0.17	0.01
<i>false</i>	0.30	0.05	0.01	0.01

Joint Probability Distribution

All relevant probabilities about a vector $\langle X_1, \dots, X_n \rangle$ of random variables can be computed from $\mathbf{P}(X_1, \dots, X_n)$

	$S = \textit{sunny}$	$S = \textit{cloudy}$	$S = \textit{rain}$	$S = \textit{snow}$	$\mathbf{P}(W)$
W	0.30	0.15	0.17	0.01	0.63
$\neg W$	0.30	0.05	0.01	0.01	0.37
$\mathbf{P}(S)$	0.60	0.20	0.18	0.02	1.00

$$P(S = \textit{rain} \wedge W) = 0.17$$

$$\begin{aligned} P(S = \textit{rain}) &= P(S = \textit{rain} \wedge W) + P(S = \textit{rain} \wedge \neg W) \\ &= 0.17 + 0.01 = 0.18 \end{aligned}$$

$$P(W) = 0.30 + 0.15 + 0.17 + 0.01 = 0.63$$

$$\begin{aligned} P(S = \textit{rain} \mid W) &= P(S = \textit{rain} \wedge W) / P(W) \\ &= 0.17 / 0.63 = 0.27 \end{aligned}$$

Joint Probability Distribution

A joint probability distribution $\mathbf{P}(X_1, \dots, X_n)$ provides complete information about the probabilities of its random variables.

However, JPD's are often hard to create (again because of incomplete knowledge of the domain).

Even when available, JPD tables are very expensive, or impossible, to store because of their size.

A JPD table for n random variables, each ranging over k distinct values, has k^n entries!

A better approach is to come up with conditional probabilities as needed and compute the others from them.

An Alternative to JPD: The Bayes Rule

Recall that for any fact A and B ,

$$P(A \wedge B) = P(A | B)P(B) = P(B | A)P(A)$$

From this we obtain the **Bayes Rule**:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)}$$

The rule is useful in practice because it is often easier to compute (or estimate) $P(A | B)$, $P(B)$, and $P(A)$ than to compute $P(B | A)$ directly

Applying the Bayes Rule

What is the probability that a patient has meningitis (M) given that he has a stiff neck (S)?

$$P(m | s) = \frac{P(s | m) P(m)}{P(s)}$$

- $P(s | m)$ is easier to estimate than $P(m | s)$ because it refers to **causal knowledge**: meningitis typically causes stiff neck
- $P(s | m)$ can be estimated from past medical cases and the knowledge about how meningitis works
- Similarly, $P(m)$ and $P(s)$ can be estimated from statistical information

Applying the Bayes Rule

The Bayes rule is helpful even in absence of (immediate) causal relationships

What is the probability that a blonde person (B) is Swedish (S)?

$$P(s | b) = \frac{P(b | s)P(s)}{P(b)}$$

All $P(b | s)$, $P(s)$, $P(b)$ are easily estimated from statistical information

$$P(b | s) \approx \frac{\# \text{ of blonde Swedish}}{\text{Swedish population}} = \frac{9}{10}$$

$$P(s) \approx \frac{\text{Swedish population}}{\text{world population}} = \dots$$

$$P(b) \approx \frac{\# \text{ of blondes}}{\text{world population}} = \dots$$

Conditional Independence

In terms of exponential explosion, conditional probabilities do not seem any better than JPD's for computing the probability of a fact, given $n > 1$ pieces of evidence

$$P(\textit{meningitis} \mid \textit{stiffNeck} \wedge \textit{nausea} \wedge \dots \wedge \textit{doubleVision})$$

However, facts do not always depend on **all** the evidence.

Example:

$$P(\textit{meningitis} \mid \textit{stiffNeck} \wedge \textit{astigmatic}) = P(\textit{meningitis} \mid \textit{stiffNeck})$$

Meningitis and *Astigmatic* are conditionally independent, given knowledge of *StiffNeck*

Conditional Probability Notation

Recall:

$$P(A \wedge B) = P(A | B)P(B) = P(B | A)P(A)$$

A general version holds for whole joint probability distributions, e.g.,

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather | Cavity) \mathbf{P}(Cavity)$$

stands for (with W for *Weather* and C for *Cavity*):

$$\begin{aligned} P(W = sunny, C = true) &= P(W = sunny | C = true) P(C = true) \\ P(W = sunny, C = false) &= P(W = sunny | C = false) P(C = false) \\ &\vdots \\ P(W = snow, C = false) &= P(W = snow | C = false) P(C = false) \end{aligned}$$

I.e., a 4×2 set of equations, **not** matrix multiplication

Conditional Probability Notation

The **chain rule** is derived by successive application of product rule:

$$\begin{aligned} & \mathbf{P}(X_1, \dots, X_n) \\ &= \mathbf{P}(X_1, \dots, X_{n-1}) \mathbf{P}(X_n | X_1, \dots, X_{n-1}) \\ &= \mathbf{P}(X_1, \dots, X_{n-2}) \mathbf{P}(X_{n-1} | X_1, \dots, X_{n-2}) \mathbf{P}(X_n | X_1, \dots, X_{n-1}) \\ &= \prod_{i=1}^n \mathbf{P}(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

Inference by enumeration

Let \mathbf{X} be all the variables. Typically, we want

the posterior joint distribution of the query variables \mathbf{Y}

given specific values \mathbf{e} for the evidence variables \mathbf{E}

Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$. Then,

$$P(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha P(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \sum_{\mathbf{h}} P(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$$

where $\alpha = P(\mathbf{E} = \mathbf{e})^{-1}$

Problems:

1. Worst-case time complexity $O(d^n)$ where d is the largest arity
2. Space complexity $O(d^n)$ to store the joint distribution
3. How to find the numbers for $O(d^n)$ entries?

Independence

A and B are independent iff

$$\mathbf{P}(A | B) = \mathbf{P}(A) \quad \text{iff} \quad \mathbf{P}(B | A) = \mathbf{P}(B) \quad \text{iff} \quad \mathbf{P}(A, B) = \mathbf{P}(A) \mathbf{P}(B)$$

Example:

$$\begin{aligned} &\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}, \textit{Weather}) \\ &= \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Weather}) \end{aligned}$$

32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do? Exploit **conditional** independence

Conditional independence

$\mathbf{P}(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

$$P(\textit{catch} \mid \textit{toothache}, \textit{cavity}) = P(\textit{catch} \mid \textit{cavity})$$

The same independence holds if I have no cavity:

$$P(\textit{catch} \mid \textit{toothache}, \neg\textit{cavity}) = P(\textit{catch} \mid \neg\textit{cavity})$$

Catch is conditionally independent on *Toothache* given *Cavity*:

$$\mathbf{P}(\textit{Catch} \mid \textit{Toothache}, \textit{Cavity}) = \mathbf{P}(\textit{Catch} \mid \textit{Cavity})$$

Conditional independence contd.

Write out full joint distribution using chain rule:

$$\begin{aligned} & \mathbf{P}(Toothache, Catch, Cavity) \\ &= \mathbf{P}(Toothache \mid Catch, Cavity) \mathbf{P}(Catch, Cavity) \\ &= \mathbf{P}(Toothache \mid Catch, Cavity) \mathbf{P}(Catch \mid Cavity) \mathbf{P}(Cavity) \\ &= \mathbf{P}(Toothache \mid Cavity) \mathbf{P}(Catch \mid Cavity) \mathbf{P}(Cavity) \end{aligned}$$

I.e., $2 + 2 + 1 = 5$ independent numbers
(equations 1 and 2 remove 2)

In most cases, conditional independence reduces the size of the representation of the JPD from exponential to linear in n

Conditional independence is our most basic and robust form of knowledge about uncertain environments

Bayes' Rule and cond. independence

$$\begin{aligned}\mathbf{P}(Cavity \mid toothache \wedge catch) \\ &= \alpha \mathbf{P}(toothache \wedge catch \mid Cavity) \mathbf{P}(Cavity) \\ &= \alpha \mathbf{P}(toothache \mid Cavity) \mathbf{P}(catch \mid Cavity) \mathbf{P}(Cavity)\end{aligned}$$

This is an example of a **naive Bayes** model:

$$\mathbf{P}(Cause, Effect_1, \dots, Effect_n) = \mathbf{P}(Cause) \prod_{i=1}^n \mathbf{P}(Effect_i \mid Cause)$$

Total number of parameters is **linear** in n

Bayesian Networks

Exploiting conditional independence information is crucial in making (automated) probabilistic reasoning feasible

Bayesian Networks are a successful example of probabilistic systems that exploit conditional independence to reason **efficiently** under uncertainty